This article has been accepted for publication in IEEE Intelligent Systems but has not yet been fully edited.
Some content may change prior to final publication.

Semantic Multi-Dimensional Scaling for Open-Domain Sentiment Analysis

Erik Cambria, *Member, IEEE,* Yangqiu Song, *Member, IEEE,*
Haixun Wang, *Member, IEEE,* and Newton Howard, *Member, IEEE*

Abstract—The ability to understand natural language text is far from being emulated in machines. One of the main hurdles to overcome is that computers lack both the common and common-sense knowledge humans normally acquire during the formative years of their lives. In order to really understand natural language, a machine should be able to grasp such kind of knowledge, rather than merely relying on the valence of keywords and word co-occurrence frequencies. In this work, the largest existing taxonomy of common knowledge is blended with a natural-language-based semantic network of common-sense knowledge, and multi-dimensional scaling is applied on the resulting knowledge base for open-domain opinion mining and sentiment analysis.

Keywords—Knowledge-Based Systems; Semantic Networks; NLP; Opinion Mining and Sentiment Analysis.

1 Introduction

THE EVER-GROWING amount of available information in the Social Web fostered the proliferation of many business and research activities around the relatively new fields of opinion mining and sentiment analysis. The automatic analysis of user generated contents such as online news, reviews, blogs, and tweets, in fact, can be extremely valuable for tasks such as mass opinion estimation, corporate reputation measurement, political orientation categorisation, stock market prediction, customer preference, and public opinion study.

Distilling useful information from such unstructured data, however, is a multi-faceted and multi-disciplinary problem, as opinions and sentiments can be expressed in a multitude of forms and combinations in which it is extremely difficult to find any kind of regular behavior. A lot of conceptual rules, in fact, govern the expression of opinions and sentiments and there exist even more clues that can convey these concepts from realisation to verbalisation in the human mind.

Most of current approaches to opinion mining and sentiment analysis rely on rather unambiguous affective keywords extracted from an existing knowledge base, e.g., WordNet [1], or from a purpose-built lexicon based on a domain-dependent corpus [2], [3], [4]. Such approaches are still far from being able to perfectly extract the conceptual and affective information associated with natural language and, hence, often fail to meet the gold standard of human annotators. Especially when dealing with social media, in fact, contents are often very diverse and noisy, and the use of a limited number of affect words or a domain-dependent training corpus is simply not enough. In order to intelligently process open-domain textual resources, computers need to be provided with both the common and common-sense knowledge humans normally acquire during the formative years of their lives, as relying just on valence of keywords and word co-occurrence frequencies does not allow a deep understanding of natural language.

In this work, ProBase [5], the largest existing taxonomy of common knowledge, is blended with ConceptNet [6], a natural-language-based semantic network of common-sense knowledge, and multi-dimensional scaling (MDS) is applied on the resulting knowledge base for sentiment analysis.

The structure of the paper is as follows: Section 2 presents related works in the field of opinion mining; Section 3 discusses how and why blending common and common-sense knowledge is important for the development of domain-independent sentiment analysis system; Section 4 explains in detail the strategies adopted to build the common and common-sense knowledge base; Section 5 illustrates how semantic MDS is employed to perform reasoning on the newly-built knowledge base; Section 6 presents the development of an opinion mining engine and its evaluation; Section 7, finally, comprises concluding remarks and future directions.

2 Related Work

Early works in the field of opinion mining and sentiment analysis aimed to classify entire documents as containing overall positive or negative polarity [7] or rating scores (e.g., 1-5 stars) of reviews [8]. These were mainly supervised approaches relying on manually-labelled samples

- *E. Cambria and N. Howard are with the Media Laboratory, MIT, 20 Ames Street, Cambridge, Massachusetts 02139–4307, USA*
  E-mail: cambria@media.mit.edu, nhmit@mit.edu
- *Y. Song is with Department of Computer Science and Engineering, Hong Kong University of Science and Technology, 1 University Ave, Hong Kong*
  E-mail: yqsong@cse.ust.hk
- *H. Wang is with Microsoft Research Asia, 5 Dan Ling Street, Beijing, 100080, P.R. China*
  E-mail: haixun.wang@microsoft.com

Digital Object Indentifier 10.1109/MIS.2012.118 0885-9000/$26.00 2012 IEEE

such as movie or product reviews, where the opinionist's overall positive or negative attitude was explicitly indicated. However, opinions and sentiments do not occur only at document-level, nor they are limited to a single valence or target. Contrary or complementary attitudes toward the same topic or multiple topics can be present across the span of a document.

Later works adopted a segment- or paragraph-level opinion analysis aiming to distinguish sentimental from non-sentimental sections, e.g., by performing a classification based on some fixed syntactic phrases likely to be used to express opinions [2] or by bootstrapping using a small set of seed opinion words and a knowledge base such as WordNet [3]. In other works, text analysis granularity has been taken down to sentence-level, e.g., by using presence of opinion-bearing lexical items (single words or n-grams) to detect subjective sentences [4] or by using semantic frames for identifying the topics (or targets) of sentiment [9].

The main aim of this work is to build possibly the most comprehensive resource of common and common-sense knowledge and apply MDS on it, in order to perform a domain-independent concept-level analysis of opinion and sentiments on the Web. A pioneering work on understanding and visualising the affective information associated to natural language text through MDS was conducted by Osgood et al. [10]. Osgood used MDS to create visualisations of affective words based on similarity ratings of the words provided to subjects from different cultures. In Osgood's work, words can be thought of as points in a multi-dimensional space and the similarity ratings represent the distances between these words. MDS projects such distances to points in a smaller dimensional space.

In this work, similarly, MDS is applied to a common and common-sense knowledge base to grasp the semantic and affective similarity between different concepts, after plotting them into a multi-dimensional vector space. Differently from Osgood's space, however, the building blocks of our vector space are not simply a limited set of similarity ratings between affect words, but rather millions of confidence scores related to pieces of common-sense knowledge linked to a hierarchy of affective domain labels. Rather than merely determined by a few human annotators and represented as a word-word matrix, in fact, our vector space is built upon a common-sense knowledge base represented as a concept-feature matrix.

3 Common and Common Sense

In standard human-to-human communication, people usually refer to existing facts and circumstances and build new useful, funny, or interesting information on the top of those. This common knowledge comprehends information usually found in news, articles, debates, lectures, etc. (factual knowledge), but also principles and definitions that can be found in collective intelligence projects such as Wikipedia[1] (vocabulary knowledge). Moreover, when people communicate with each other, they rely on similar background knowledge, e.g., the way objects relate to each other in the world, people's goals in their daily lives, and the emotional content of events or situations. This taken-for-granted information is what is termed common sense – obvious things people normally know and usually leave unstated.

3.1 Common Knowledge Base

Attempts to build a common knowledge base are countless and comprehend both resources crafted by human experts or community efforts, such as WordNet, with its 25,000 synsets, or Freebase [11], a social database of 1,450 concepts, and automatically-built knowledge bases, such as YAGO [12], a semantic knowledge base of 149,162 instances derived from Wikipedia, WordNet, and GeoNames[2], and ProBase.

ProBase contains about 12 million concepts learned iteratively from 1.68 billion web pages in Bing[3] web repository. The taxonomy is probabilistic, which means every claim in ProBase is associated with some probabilities that model the claim's correctness, ambiguity, and other characteristics. The probabilities are derived from evidences found in web data, search log data, and other available data. The core taxonomy consists of the "IsA" relationships extracted by using syntactic patterns. For example, a segment like "artists such as Pablo Picasso" can be considered as a piece of evidence for the claim that 'pablo picasso' is an instance of the concept 'artist'.

3.2 Common-Sense Knowledge Base

One of the biggest projects aiming to build a comprehensive common-sense knowledge base is Cyc [13]. Cyc, however, requires the involvement of experts working on specific languages and contains just 120,000 concepts, as the knoweldge engineering is labor-intensive and time-consuming. A more recent and scalable project is Open Mind Common Sense (OMCS), which has been collecting pieces of knowledge from volunteers on the Internet since 1999, by enabling the general public to enter common sense into the system with no special training or knowledge of computer science. OMCS exploits these pieces of common-sense knowledge to automatically build ConceptNet, a semantic network of 173,398 nodes (upon which many other common-sense resources, e.g., SenticNet[4], are built).

WordNet contains very detailed descriptions of every word's various senses, but it does not include enough general Web information. ProBase, which provides more concepts, includes pieces of knowledge that match general distribution of human knowledge. ConceptNet, in

[1] http://wikipedia.org
[2] http://geonames.org
[3] http://bing.com
[4] http://sentic.net

Digital Object Indentifier 10.1109/MIS.2012.118 0885-9000/$26.00 2012 IEEE

turn, contains implicit knowledge that people rarely mention on the Web, which is good complementary material to Probase.

4 Building the Knowledge Base

In [14], Probase IsA relationships were exploited to build a semantic network, termed Isanette (IsA net), representing hyponym-hypernym common knowledge as a matrix having instances (e.g., 'pablo picasso') as rows and concepts (e.g., 'artist') as columns. In this work, an extended version of Probase is used and the new Isanette matrix is 4,622,119 × 2,524,453. Because Isanette is a very large and fat matrix that contains noise and multiple forms, it is firstly cleaned by applying different natural language processing (NLP) and MDS techniques (Section 4.1). Secondly, Isanette's consistency is enhanced (and its sparseness further reduced) by adding complementary common-sense knowledge (Section 4.2).

4.1 Cleaning Isanette

Isanette is built out of about 40 million IsA triples extracted with the form <instance, concept, confidence score>. Before generating the matrix from these statements, however, two main issues need to be solved: multiple concept forms and low connectivity.

The first issue is addressed by exploiting both word similarity and MDS. The concept 'barack obama', for example, appears in the triples in many different forms such as 'president obama', 'mr barack obama', 'president barack obama', etc. Trying to disambiguate this kind of instances *a priori*, by simply using word similarity, could be dangerous as concepts like 'buy christmas present' and 'present christmas event' have very different meanings, although they have high word similarity. Hence, an *a posteriori* concept deduplication is performed by exploiting concept semantic relatedness, after Isanette is built. That is, concepts with high word similarity are merged together just if they are close enough to each other in the vector space generated from Isanette.

The second issue is addressed by discarding hapax legomena, that is, instances/concepts with singular out-/in-degree. If MDS is to be applied in order to find similar patterns, in fact, Isanette is to be as less sparse as possible. In this work, not only hapax legomena are discarded, but also other long-tail concepts, in order to heavily enhance Isanette's graph connectivity. In particular, a trial-and-error approach is used to find that the best trade-off between size and sparseness is achieved by setting the minimum node connectivity equal to 10. This cut-off operation leaves out almost 40% of nodes and makes Isanette a strongly connected core. Moreover, MDS is exploited to infer negative evidence such as 'carbonara' is not a kind of 'fuel' or 'alitalia' is not a 'country', which is very useful to further reduce Isanette's sparseness and improve reasoning algorithms.

4.2 Blending Isanette

As a subsumption common knowledge base, Isanette lacks information like a 'dog' is a 'best friend' (rather than simply an 'animal') or a 'rose' is a kind of 'meaningful gift' (rather than simply a kind of 'flower'), i.e., common sense that is not usually stated in web pages (or at least not that often to be extracted by Hearst patterns with a high-enough confidence score). To overcome this problem, Isanette is enriched with complementary hyponym-hypernym common-sense knowledge from ConceptNet. In particular, all the assertions involving IsA relationships with a non-null confidence score, such as "dog is man's best friend" or "a birthday party is a special occasion", are extracted from the Open Mind corpus. Such assertions are exploited to generate a directed graph of about 15,000 nodes (interconnected by IsA edges), representing subsumption common-sense knowledge.

To merge this subsumption common-sense knowledge base with Isanette, the blending technique [15] is employed. Blending is a technique that performs inference over multiple sources of data simultaneously, taking advantage of the overlap between them. It combines two sparse matrices linearly into a single matrix, in which the information between the two initial sources is shared. This alignment operation yields IsaCore[5], a new strongly-connected core (hereafter referred as C, for the sake of simplicity) in which common and common-sense knowledge coexist, i.e., a matrix 500,000 × 300,000 whose rows are instances (e.g., 'birthday party' and 'china'), whose columns are concepts (e.g., 'special occasion' and 'country'), and whose values indicate truth values of assertions.

5 Reasoning on the Knowledge Base

In this section, MDS is applied to build a vector space representation of the instance-concept relationship matrix (Section 5.1). A semi-supervised learning algorithm is then employed to further discriminate affective information (Section 5.2), and a partitioning clustering technique is used to segment the reduced space into conceptual classes (Section 5.3).

5.1 Semantic Multi-Dimensional Scaling

In order to more-compactly represent the information contained in $C \in \mathcal{R}^{m \times n}$ and encode the latent semantics between its instances, a multi-dimensional vector space representation is built by applying truncated singular value decomposition (SVD). The resulting lower-dimensional space represents the best approximation of C, in fact:

$$\begin{aligned} \min_{\tilde{C}|rank(\tilde{C})=d} |C - \tilde{C}| &= \min_{\tilde{C}|rank(\tilde{C})=d} |\Sigma - U_d^T \tilde{C} V_d| \\ &= \min_{\tilde{C}|rank(\tilde{C})=d} |\Sigma - S_d| \end{aligned}$$

[5] http://sentic.net/isacore.zip

where C has the form $C = U\Sigma V^T$, $\tilde{C}$ has the form $\tilde{C} = U_d S_d V_d^T$ ($U_d \in \mathcal{R}^{m\times d}$, $V_d \in \mathcal{R}^{n\times d}$, and $S_d \in \mathcal{R}^{d\times d}$ is diagonal matrix), and d is the lower dimension of the latent semantic space. From the rank constraint, i.e., S_d has d non-zero diagonal entries, the minimum of the above statement is obtained as follows:

$$\min_{\tilde{C}|rank(\tilde{C})=d} |\Sigma - S_d| = \min_{s_i} \sqrt{\sum_{i=1}^{n} (\sigma_i - s_i)^2} =$$

$$= \min_{s_i} \sqrt{\sum_{i=1}^{d} (\sigma_i - s_i)^2 + \sum_{i=d+1}^{n} \sigma_i^2} = \sqrt{\sum_{i=d+1}^{n} \sigma_i^2}$$

Therefore, $\tilde{C}$ of rank d is the best approximation of C in the Frobenius-norm sense when $\sigma_i = s_i$ $(i = 1, ..., d)$ and the corresponding singular vectors are the same as those of C. If all but the first d principal components are discarded and $\tilde{C}_U = U_d\, S_d$ is considered, a space in which common and common-sense instances are represented by vectors of d coordinates is obtained.

These coordinates can be seen as describing instances in terms of 'eigenconcepts' that form the axes of the vector space, i.e., its basis $e = (e^{(1)}$,... $e^{(d)})^T$. A trial-and-error approach is used to find that the best compromise is achieved when d assumes values around 500. Such a 500-dimensional vector space can be used for making analogies (given a specific instance, find the instances most semantically related to it), for making comparisons (given two instances, infer their degree of semantic relatedness), and for classification purposes (given a specific instance, assign it to a predefined cluster).

5.2 Semi-Supervised Affective Propagation

After applying SVD in Section 5.1, the obtained $\tilde{C}_U$ does not lead to meaningful affective relatedness results, as the vector space represents semantic relatedness of instances according to IsA relationships. Affectively opposite instances such as 'smile' and 'cry', in fact, are likely to be found close to each other in $\tilde{C}_U$ because they both relate to emotions. Hence, in order to build an appropriate space that is both semantic and sentiment preserving, a semi-supervised linear discriminant analysis (LDA) algorithm is adopted.

Differently from other classifiers, semi-supervised LDA can incorporate both supervised (affective keywords) and unsupervised (non-affective keywords) information in such a way that a proper semantic space that reflects the sentiment information is obtained. In this work, LDA is also preferred to other classifiers for its analytical simplicity and computational efficiency. More in-depth motivations for the choice of LDA can be found in [16].

In order to infer affective information from natural language and use it for tasks such as emotional labelling and opinion polarity detection, existing approaches rely on a relatively small set of affective words extracted from manually-labelled lexicons, e.g., WordNet, and a few emotional labels, e.g., Ekman's six universal emotions. Since such categorical approaches classify emotions using a list of labels, they usually fail to describe the complex range of emotions that can occur in daily communication.

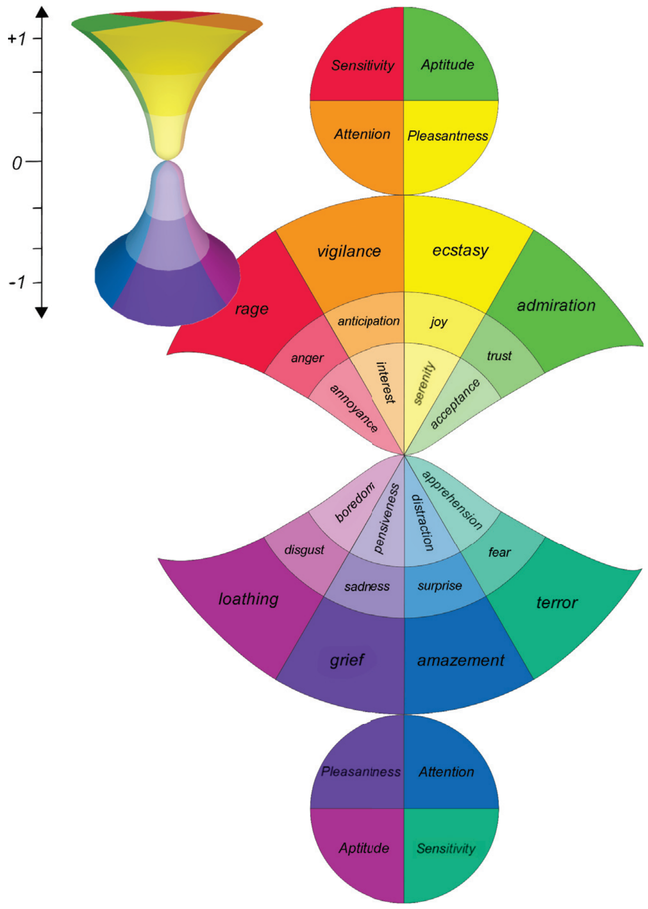

Fig. 1. Hourglass of Emotions

To overcome this problem, the Hourglass of Emotions [17] categorisation model (Fig. 1) is adopted. Because such a biologically-inspired and psychologically-motivated model goes beyond mere categorical and dimensional approaches, it can potentially describe any human emotion in terms of four independent but concomitant dimensions.

Given a set of affective labels and a large amount of unlabelled instances in C, the between-class scatter is to be maximized and the within-class scatter of expressly affective instances (from the Hourglass model) is to be minimized, as well as the semantic relatedness of all the other instances simultaneously is to be kept.

Each instance is denoted as $e_i \in \mathcal{R}^d$, which is a d-dimensional vector after processed with SVD. For each expressly affective instance, there is a label $y_i \in \{1, \ldots, q\}$, where q is the number of sentiment classes. Then, the between-class scatter and the within-class scatter matrices are defined as follow:

$$S_w = \sum_{j=1}^{q} \sum_{i=1}^{l_j} (e_i - \mu_j)(e_i - \mu_j)^T$$

$$S_b = \sum_{j=1}^{q} l_j(\mu_j - \mu)(\mu_j - \mu)^T$$

where $\mu_j = \frac{1}{l_j} \sum_{i=1}^{l_j} e_i$ $(j = 1, 2, ..., q)$ is the mean of the samples in class j, l_j is the number of affective instances in class j and $\mu = \frac{1}{l} \sum_{i=1}^{l} e_i$ is the mean of all the labelled samples. A total scatter matrix on all the instances in C is also defined:

$$S_t = \sum_{i=1}^{m} (e_i - \mu_m)(e_i - \mu_m)^T$$

where m is the total number of instances in C and μ_m is the mean of all the instances. Our objective is then to find a projection matrix W to project the semantic space to a lower-dimensional space, which is more affectively discriminative:

$$W^* = \arg \max_{W \in \mathcal{R}^{d \times d'}} \frac{|W^T S_b W|}{|W^T(S_w + \lambda_1 S_t + \lambda_2 I)W|}$$

where I is identity matrix, and λ_1 and λ_2 are control parameters, obtained through a grid search, for balancing the trade-off between sentiment discriminant and semantic regularisations. The optimal solution is given by:

$$(S_w + \lambda_1 S_t + \lambda_2 I)w_j^* = \eta_j S_b w_j^* \quad j = 1, ..., d'$$

where w_j^* $(j = 1, ..., d')$ are the eigenvectors corresponding to the d' largest eigenvalues of $(S_w + \lambda_1 S_t + \lambda_2 I)^{-1} S_b$. Here $d' = q - 1$ is selected, where q is the total emotion number. After the projection, the new space preserves both semantic and sentiment property based on the instance-concept relationships and affective labels.

5.3 Semantic Clustering

In order to perform concept-level topic-spotting in natural language opinions, different degrees of membership to different classes are to be assigned to each instance. To this end, $\tilde{C}_U$ is clustered into k distinct categories represented by Isanette's hub concepts, that is, the top 5,000 concepts with highest in-degree in Isanette.

A sentic medoids [18] approach is employed. Differently from the k-means algorithm (which does not pose constraints on centroids), sentic medoids do assume that centroids must coincide with k observed points, which allows to better cluster a vector space of common-sense knowledge [18]. The sentic medoids approach is similar to the partitioning around medoids (PAM) algorithm, which determines a medoid for each cluster selecting the most centrally located centroid within that cluster. Unlike other PAM techniques, however, the sentic medoids algorithm runs similarly to k-means and, hence, requires a significantly reduced computational time.

Generally, the initialisation of clusters for clustering algorithms is a problematic task as the process often risks to get stuck into local optimum points, depending on the initial choice of centroids. For this study, however, the most representative (highest confidence score) instances of Isanette's hub concepts are used as initial centroids. For this reason, what is usually seen as a limitation of the algorithm can be seen as advantage for this study, since what is being sought is not the 5,000 centroids leading to the best 5,000 clusters, but indeed the 5,000 centroids identifying the top 5,000 hub concepts (i.e., the centroids should not be 'too far' from the most representative instances of these concepts). Therefore, given that the distance between two points in the space is defined as $D(e_i, e_j) = \sqrt{\sum_{s=1}^{d'} \left(e_i^{(s)} - e_j^{(s)}\right)^2}$, the adopted algorithm can be summarised as follows:

1) Each centroid $\bar{e}_i \in \mathbb{R}^{d'}$ $(i = 1, 2, ..., k)$ is set as one of the k most representative instances of the top hub concepts;
2) Assign each instance e_j to a cluster $\bar{e}_i$ if $D(e_j, \bar{e}_i) \leq D(e_j, \bar{e}_{i'})$ where $i(i') = 1, 2, ..., k$;
3) Find a new centroid $\bar{e}_i$ for each cluster c so that $\sum_{j \in Cluster\ c} D(e_j, \bar{e}_i) \leq \sum_{j \in Cluster\ c} D(e_j, \bar{e}_{i'})$;
4) Repeat step 2 and 3 until no changes on centroids are observed.

6 Exploiting the Knowledge Base

In order to assess the accuracy of IsaCore, an opinion mining engine able to infer both the conceptual and affective information associated with natural language text was developed. Such an engine consists of four main components: a pre-processing module, which performs a first skim of the opinion; a semantic parser, whose aim is to extract concepts from the opinionated text; a target spotting module, which identifies opinion targets; finally, an affect interpreter, for emotion recognition and polarity detection.

The pre-processing module firstly interprets special punctuation, complete upper-case words, cross-linguistic onomatopoeias, exclamation words, negations, degree adverbs, and emoticons. Secondly, it converts text to lower-case and, after lemmatising it, splits the opinion into single clauses according to grammatical conjunctions and punctuation.

Then, the semantic parser deconstructs text into small bags of concepts (SBoCs) using a lexicon based on sequences of lexemes that represent multiple-word concepts extracted from ConceptNet and Isanette. These n-grams are not used blindly as fixed word patterns but exploited as reference for the module, in order to extract multiple-word concepts from information-rich sentences. So, differently from other shallow parsers, the module can recognise complex concepts also when these are interspersed with adjective and adverbs, e.g., the concept

Digital Object Indentifier 10.1109/MIS.2012.118 0885-9000/$26.00 2012 IEEE

'buy christmas present' in the sentence "I bought a lot of very nice Christmas presents".

The target spotting module aims to individuate one or more opinion targets, such as people, places, and events, from the input concepts. This is done by projecting the concepts of each SBoC into $\widetilde{C}_U$, clustered according to Isanette's hub concepts. The categorisation does not consist in simply labelling each concept, but also in assigning a confidence score to each category label, which is directly proportional to the value of belonging (dot product) to a specific conceptual cluster.

The affect interpreter, similarly, projects the concepts of each SBoC into $\widetilde{C}_U$, clustered according to the Hourglass labels, and, hence, calculates polarity in terms of the Hourglass dimensions (Pleasantness, Attention, Sensitivity, and Aptitude) according to the formula [17]:

$$p = \sum_{i=1}^{N} \frac{Plsnt(c_i) + |Attnt(c_i)| - |Snst(c_i)| + Aptit(c_i)}{3N}$$

where c_i is an input concept, N the size of the SBoC, and 3 the normalisation factor.

In order to evaluate the different facets of the opinion mining engine from different perspectives, three different resources, namely a Twitter[6] hashtag repository, a LiveJournal[7] database, and a PatientOpinion[8] dataset are used, and results obtained using WordNet, ConceptNet, and Isanette are compared.

The first resource is a collection of 3,000 tweets crawled from Bing web repository by exploiting Twitter hashtags as category labels, which is useful to test the engine's target spotting performances. In particular, hashtags about electronics (e.g., IPhone, XBox, Android, and Wii), companies (e.g., Apple, Microsoft, and Google), countries, cities, operative systems, and cars are selected.

The second resource is a 5,000-blogpost database extracted from LiveJournal, a virtual community of more than 23 millions users who keep a blog, journal, or diary. An interesting feature of this website is that bloggers are allowed to label their posts with a mood tag, by choosing from predefined mood themes or by creating new ones. Since the indication of mood tags is optional, posts are likely to reflect the authors' true mood.

The third resource, finally, is a dataset obtained from PatientOpinion, a social enterprise pioneering an on-line feedback service for users of the UK national health service to enable people to share their experience of local health services. It is a manually-tagged dataset of 2,000 patient opinions that associates to each post a category (namely, clinical service, communication, food, parking, staff, and timeliness) and a positive or negative polarity.

In order to assess the accuracy of IsaCore, a comparison study was carried out by replacing it with state-of-the-art knowledge bases in the opinion mining engine. In particular, WordNet, ConceptNet, and Isanette were firstly swapped with IsaCore to compare topic spotting performance and emotion recognition capabilities of the engine on the Twitter hashtag repository and on the LiveJournal database, respectively. Secondly, the same evaluation process was repeated to concurrently assess the engine's topic spotting and polarity detection capabilities on the PatientOpinion dataset.

As for the Twitter evaluation, results show that Isanette and IsaCore perform significantly better than WordNet and ConceptNet, as these lack factual knowledge concepts such as Wii or Ford Focus (see Table 1, rows 1–6). Isanette's and IsaCore's topic spotting precision, on the other hand, is comparable as Isanette hyponym-hypernym common knowledge is enough for this kind of task. It actually even outperforms IsaCore sometimes as this contains just a subset of Isanette instances (hub instances) and common-sense knowledge does not play a key role in this type of classification.

As for the LiveJournal evaluation, the capability of the software engine to properly categorise antithetical affective pairs from the Hourglass model (namely joy-sadness, anticipation-surprise, anger-fear, and trust-disgust) was evaluated. Results show that, in this case, Isanette is consistently outperformed by IsaCore, as it is based on semantic, rather than affective, relatedness of concepts (F-measure values are reported in Table 1, rows 7–10). In Isanette vector space representation, in fact, instances like 'joy', 'surprise', and 'anger' are all close to each other, although they convey different affective valence, for being associated with the same hyponym-hypernym relationships.

As for the PatientOpinion evaluation, finally, IsaCore turns out to be the best choice as it represents the best trade-off between common and common-sense knowledge, which is particularly needed when aiming to infer both the conceptual and affective information associated with text (F-measure values are reported in Table 1, rows 11–16). As also shown by previous experiments, in fact, common knowledge is particularly functional for tasks such as open-domain text auto-categorisation, while common-sense knowledge is notably useful for natural language understanding and inference of implicit meaning underpinning words.

7 Conclusion

In this work, common and common-sense knowledge were blended together in order to build a comprehensive resource that can be seen as an attempt to emulate how tacit and explicit knowledge is organised in human mind, and how this can be exploited to perform reasoning within natural language tasks such as opinion mining and sentiment analysis. It is usually hard to take advantage of a knowledge base in systems different from the one the resource was conceived for. Indeed, its underlying symbolic framework and content, whilst being very efficient for its original purpose, are not flexible enough to be fruitfully exported and embedded in any application.

[6] http://twitter.com
[7] http://livejournal.com
[8] http://patientopinion.org.uk

TABLE 1
Precision values relative to Twitter evaluation (rows 1–6) and F-measure values relative to LiveJournal (rows 7–10) and PatientOpinion evaluation (rows 11–16).

	WordNet	ConceptNet	Isanette	IsaCore
electronics	34.5%	45.3%	79.1%	79.2%
companies	26.4%	51.0%	82.3%	82.3%
countries	38.2%	65.4%	85.2%	84.9%
cities	25.3%	59.3%	80.4%	81.8%
operative systems	37.3%	51.4%	77.8%	75.6%
cars	13.1%	22.2%	76.5%	76.7%
joy-sadness	47.5%	55.1%	75.5%	81.8%
anticipation-surprise	30.2%	41.4%	62.3%	73.0%
anger-fear	43.3%	49.0%	60.6%	71.6%
trust-disgust	27.3%	39.5%	58.8%	69.9%
clinical service	35.1%	49.5%	78.3%	82.9%
communication	41.0%	50.4%	71.6%	79.7%
food	39.3%	45.4%	65.9%	81.6%
parking	47.3%	51.6%	73.4%	77.8%
staff	32.9%	37.2%	69.8%	73.9%
timeliness	44.0%	50.4%	62.8%	80.8%

IsaCore is different as it is an open-domain resource and it exploits reasoning techniques able to infer general conceptual and affective information, which can be used for many different tasks such as opinion mining, affect recognition, text auto-categorisation, etc. Whilst this study has shown encouraging results, further research studies are now planned to investigate if a better trade-off between size and sparseness of IsaCore can be found. At the same time, new semantic MDS techniques are to be explored to perform reasoning on the knowledge base.

REFERENCES

[1] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[2] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *ACL*, Philadelphia, 2002, pp. 417–424.

[3] J. Kamps, M. Marx, R. Mokken, and M. de Rijke, "Using WordNet to measure semantic orientation of adjectives," in *LREC*, Lisbon, 2004, pp. 1115–1118.

[4] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *EMNLP*, Sapporo, 2003, pp. 105–112.

[5] W. Wu, H. Li, H. Wang, and K. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *SIGMOD*. Scottsdale, 2012, pp. 481–492.

[6] R. Speer and C. Havasi, "ConceptNet 5: A large semantic network for relational knowledge," in *Theory and Applications of Natural Language Processing*, E. Hovy, M. Johnson, and G. Hirst, Eds. Springer, 2012, ch. 6.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *EMNLP*, Philadelphia, 2002, pp. 79–86.

[8] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *ACL*, Ann Arbor, 2005, pp. 115–124.

[9] S. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," in *Workshop on Sentiment and Subjectivity in Text*, Sydney, 2006.

[10] C. Osgood, W. May, and M. Miron, *Cross-Cultural Universals of Affective Meaning*. Univ. of Illinois Press, 1975.

[11] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *SIGMOD*, Vancouver, 2008, pp. 1247–1250.

[12] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, Banff, 2007, pp. 697–706.

[13] D. Lenat and R. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Boston: Addison-Wesley, 1989.

[14] E. Cambria, Y. Song, H. Wang, and A. Hussain, "Isanette: A common and common sense knowledge base for opinion mining," in *ICDM*, Vancouver, 2011, pp. 315–322.

[15] C. Havasi, R. Speer, J. Pustejovsky, and H. Lieberman, "Digital intuition: Applying common sense using dimensionality reduction," *IEEE Intelligent Systems*, vol. 24, no. 4, pp. 24–35, 2009.

[16] Y. Song, F. Nie, C. Zhang, and S. Xiang, "A unified framework for semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 41, no. 9, pp. 2789–2799, 09 2008.

[17] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioral Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Muller, Eds. Berlin Heidelberg: Springer, 2012, vol. 7403, pp. 144–157.

[18] E. Cambria, T. Mazzocco, A. Hussain, and C. Eckl, "Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space," in *Advances in Neural Networks*, ser. Lecture Notes in Computer Science, D. Liu, H. Zhang, M. Polycarpou, C. Alippi, and H. He, Eds., vol. 6677. Berlin: Springer-Verlag, 2011, pp. 601–610.

Erik Cambria is an associate researcher at the Massachusetts Institute of Technology (Synthetic Intelligence Project, Media Lab). His interests include AI, Semantic Web, KR, NLP, opinion mining and sentiment analysis, affective and cognitive modeling, HCI, and e-health. He is chair of several international conferences, e.g., BICS, symposia, e.g., ELM, and workshops series, e.g., ICDM SENTIRE and KDD WISDOM. He is also editorial board of Springer Cognitive Computation and the Brain Sciences Journal.

Yangqiu Song received his BEng in July, 2003 and his PhD degree in January, 2009, from the Department of Automation, Tsinghua University, China. He is currently a post-doctoral researcher at Hong Kong University of Science and Technology. Before that, he was an associate researcher at Microsoft Research Asia, and he was a researcher at IBM Research - China. His research interests include machine learning algorithms with applications to knowledge engineering, information retrieval and visualisation.

Haixun Wang is a senior researcher at Microsoft Research Asia, where he manages the group of Data Management, Analytics, and Services. Before joining Microsoft, he had been a research staff member at IBM T. J. Watson Research Center for 9 years. He has published more than 120 research papers in referred international journals and conference proceedings. He is on the editorial board of Distributed and Parallel Databases (DAPD), IEEE Transactions of Knowledge and Data Engineering (TKDE), Knowledge and Information System (KAIS), and Journal of Computer Science and Technology (JCST).

Digital Object Indentifier 10.1109/MIS.2012.118 0885-9000/$26.00 2012 IEEE

Newton Howard is one of the directors of the Synthetic Intelligence Project and a resident scientist at the Massachusetts Institute of Technology. He has served as the Chairman of the Center for Advanced Defense Studies (CADS), the leading Washington, D.C, National Security Group and is currently its board director. He is a national security advisor to several U.S. Government organizations and works with multidisciplinary teams of physicists, chemists, biologists, brain scientists, computer scientists, and engineers to reach a deeper understanding of the brain.

Digital Object Indentifier 10.1109/MIS.2012.118 0885-9000/$26.00 2012 IEEE